

# Как собрать огромный датасет и не потратить годовой бюджет маленькой страны

Карина Кванчиани / CV Engineer / SberDevices  
Александр Капитанов / Team Lead CV RnD / SberDevices



# **Как собрать огромный датасет и не потратить годовой бюджет маленькой страны**

Карина Кванчiani / CV Engineer / SberDevices  
Александр Капитанов / Team Lead CV RnD / SberDevices

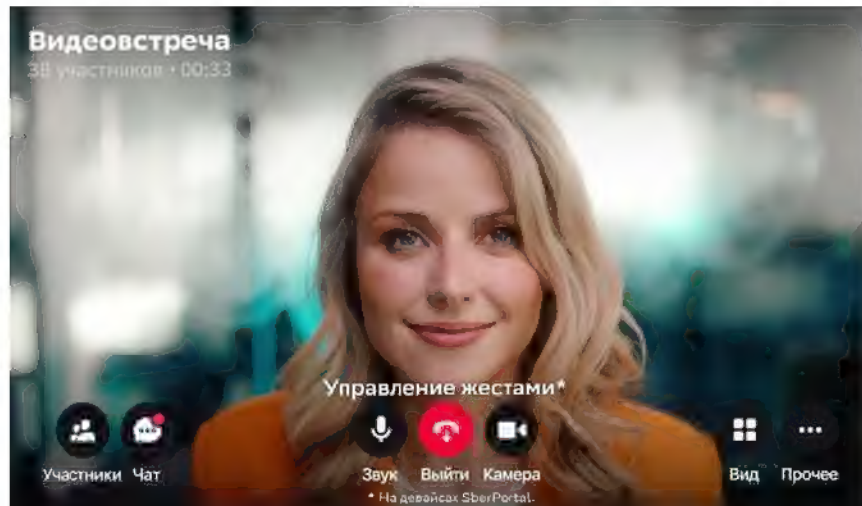
# Девайсы и Jazz

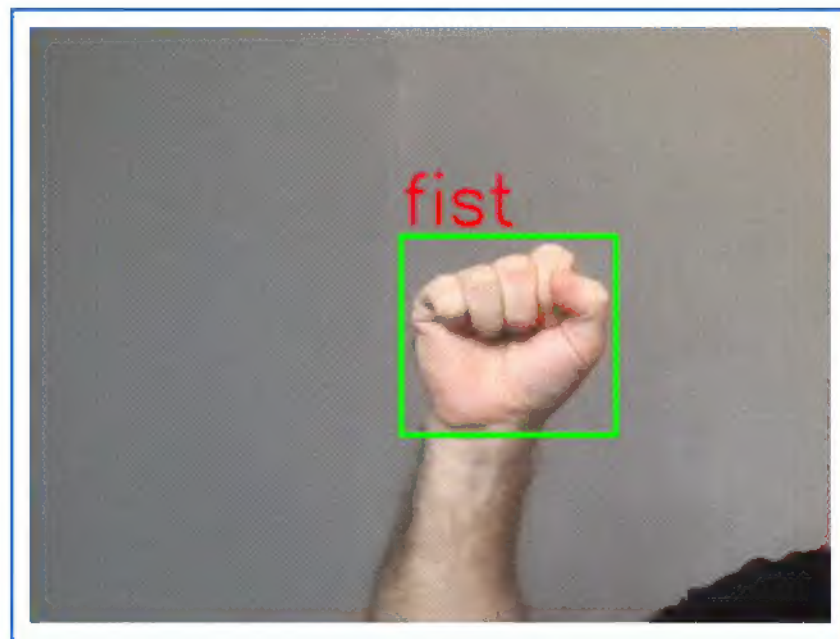


Салют



Jazz by Sber





# Распознавание жестов и данные





# HaGRID

HAnd Gesture Recongition Image Dataset



**552.992**  
изображений  
в FullHD

**18** классов  
жестов

дополнительный  
класс "не жест"

**34,730**  
различных субъектов

**>= 34,730** 🤘  
различных сцен

**от 18 до 65**  
возраст субъектов

**ж / м = 27 / 20**  
пол субъектов

искусственный и  
естественный свет +  
**экстремальные условия**

**от 0.5 до 4 метров**  
расстояние  
субъектов до камеры

## План доклада

- Краудсорсинг-платформы
- Пайплайн сбора и разметки данных
- Минимизация затрат и повышение качества разметки
- Публикация в open-source
- Автоматизация пайплайна

# Краудсорсинг-платформы



Yandex.Toloka

готовый продукт  
лучше интерфейс  
больше разметчиков



ABC Elementary

есть модератор  
"идеальные" разметчики  
кастомные проекты



# Краудсорсинг-платформы

проекты, пулы и как это все работает



# Пайплайн сбора и разметки данных

Yandex.Toloka  
сбор изображений

автоматизировано :)  
удаляем дубликаты  
& фото с низким  
разрешением

Yandex.Toloka  
валидация

Yandex.Toloka &  
ABC Elementary  
разметка  
боксами

валидация  
разметки  
Yandex.Toloka &  
ABC Elementary

HaGRID  
✌️

мы делаем  
здесь amazing  
customer  
service



# Пайплайн сбора и разметки данных

[Yandex.Toloka](#)

сбор изображений

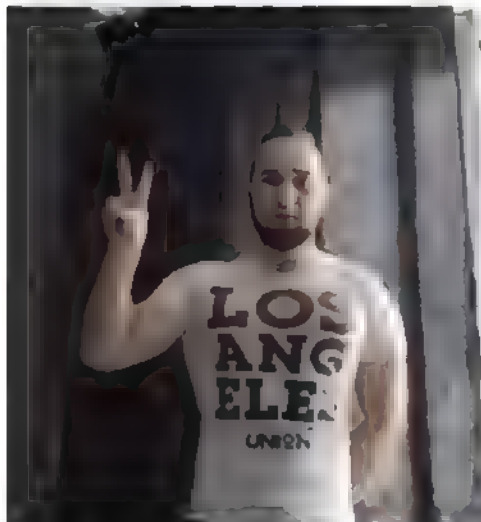
# Пайплайн сбора и разметки данных

## сбор изображений

Ваше фото с жестом

Жест виктори (указательный и средний палец образуют знак V)

Шаблон к заданию



Ваше фото

 Сделать фото

**Задача:**

показать указанный в задании жест и сделать фото себя

**Условия:**

рука с жестом должна быть в кадре полностью

фото должно быть сделано с расстояния от 1 до 4 метров

# Пайплайн сбора и разметки данных

## сбор изображений

### Соотношение скорости и качества

Обратите внимание, чем меньше исполнителей выбрано, тем ниже скорость выполнения пула

[Подробнее](#)

ТОП % ☐ Онлайн

Пул будет доступен указанному проценту исполнителей с лучшим рейтингом

10809

Скорость

1080

Качество

Мы отобрали **80%** лучших исполнителей

Задание доступно **8647** активным исполнителям

# Пайплайн сбора и разметки данных сбор изображений

## Цена

Цена за страницу заданий, \$ \* ?

Рекомендованная цена для простых заданий — 0,02 \$

Рекомендованная цена для сложных заданий — 0,05 \$

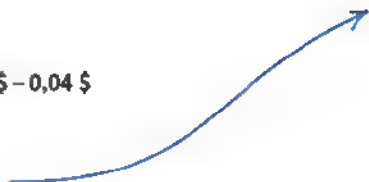
Динамическое ценообразование за страницу заданий

для навыка Запись статичных жестов

0.01



0,03 \$ – 0,04 \$



## Динамическое ценообразование и перекрытие

### Динамическое ценообразование

Укажите диапазон цен в зависимости от навыка, чтобы мотивировать исполнителей с более высоким уровнем навыка

☒ Использовать динамическое ценообразование ?

Навык \*

Запись статичных жестов



Выбрать свой навык ▾

Список навыков: Запись статичных жестов, Запись динамичных жестов, Запись статичных жестов с разметкой, Запись динамичных жестов с разметкой, Запись статичных жестов с разметкой и т.д.

Значения навыка \*

Цена за страницу заданий, \$ \*

от 0 до 95



0.03



от 96 до 100



0.04



укажите ставку для каждого уровня навыка от 0 до 100 К исполнителям у которых нет навыка не показывать задания

# Пайплайн сбора и разметки данных

## сбор изображений

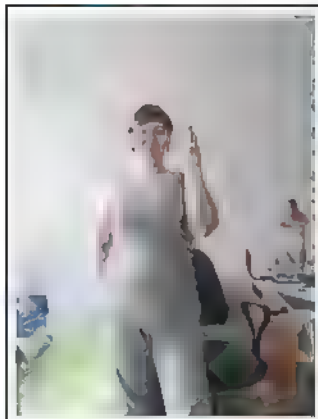
Имя задачи	Приоритет	Прогресс	Статус	Запущен	Будет завершён
CAL [RNDCV] [Alexander Kapitanov]	0	8005* 10000	Идёт разметка	18 октября 2022 г. 9:22	2 часа
LIKE [RNDCV] [Alexander Kapitanov]	0	6143* 8000	Идёт разметка...	18 октября 2022 г. 9:21	час
DISLIKE [RNDCV] [Alexander Kapitanov]	0	6134* 8000	Идёт разметка...	18 октября 2022 г. 9:21	2 часа
FAST [RNDCV] [Alexander Kapitanov]	0	6785* 8000	Идёт разметка	18 октября 2022 г. 9:21	час
STOP [RNDCV] [Alexander Kapitanov]	0	6645* 9000	Идёт разметка...	18 октября 2022 г., 22:56	11 часов
STOP_INVERTED [RNDCV] [Alexander Kapitanov]	0	1765* 2000	Идёт разметка	13 октября 2022 г. 13:20	20 часов
THREE [RNDCV] [Alexander Kapitanov]	0	6273* 8000	Идёт разметка...	18 октября 2022 г., 9:21	5 минут
TWO_UP [RNDCV] [Alexander Kapitanov]	0	712* 1000	Идёт разметка	18 октября 2022 г. 21:58	8 часов
THREE2 [RNDCV] [Alexander Kapitanov]	0	681* 1000	Идёт разметка...	18 октября 2022 г., 22:22	9 часов
FOUR [RNDCV] [Alexander Kapitanov]	0	11000 11000	Размечен	9 октября 2022 г., 19:28	—
PALM [RNDCV] [Alexander Kapitanov]	0	11000 11000	Размечен	28 сентября 2022 г. 15:48	
MUTE [RNDCV] [Alexander Kapitanov]	0	11000 11000	Размечен	7 октября 2022 г. 17:05	



# Пайплайн сбора и разметки данных

## сбор изображений

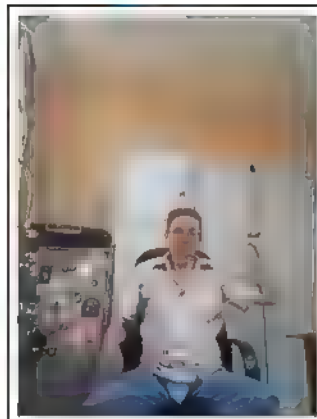
 two up



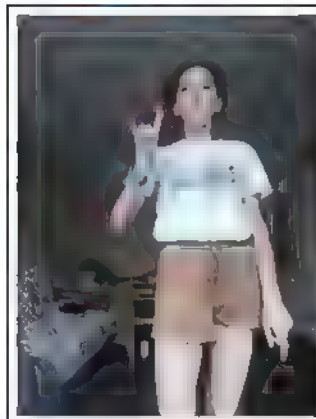
 ok



 fist



 rock

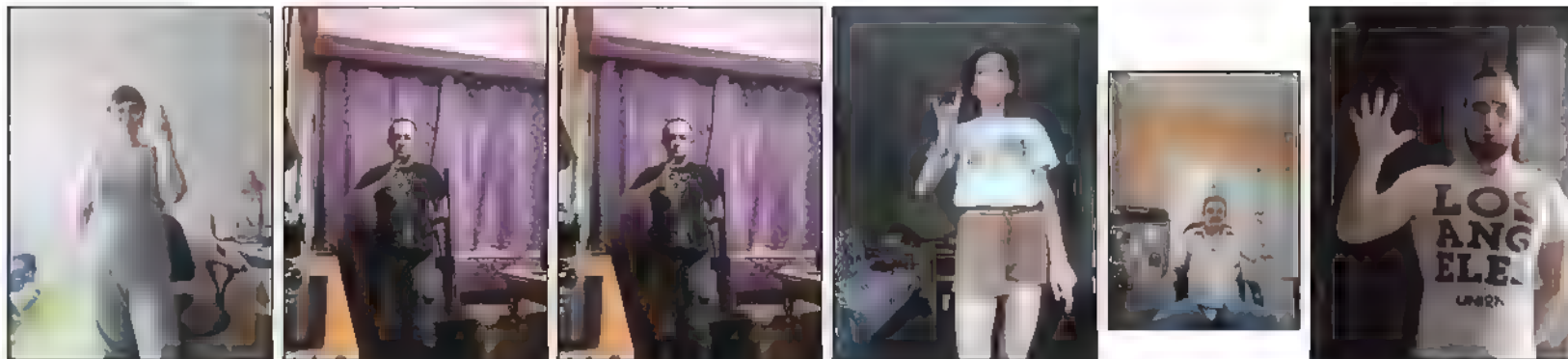


## Пайплайн сбора и разметки данных

автоматизировано :)  
удаляем дубликаты  
& фото с низким  
разрешением

# Пайплайн сбора и разметки данных

удаляем дубликаты & фото с низким разрешением

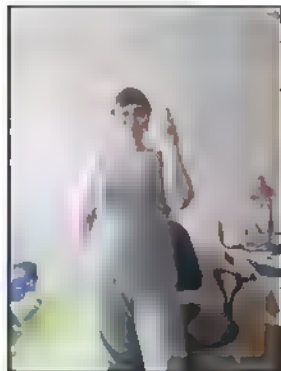


## Пайплайн сбора и разметки данных

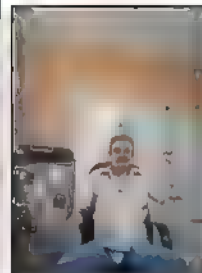
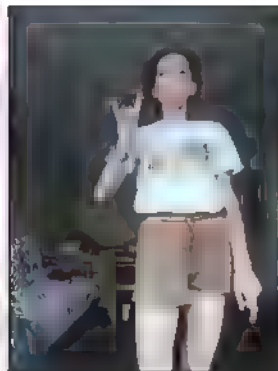
удаляем дубликаты & фото с низким разрешением



шаблон



дубликат

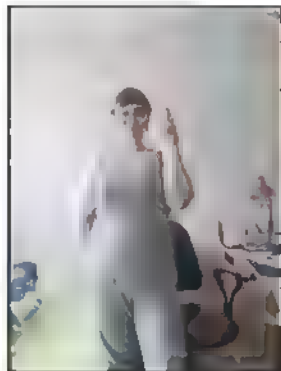


низкое разрешение



## Пайплайн сбора и разметки данных

удаляем дубликаты & фото с низким разрешением



# Пайплайн сбора и разметки данных уровни недобросовестных исполнителей

O IVI

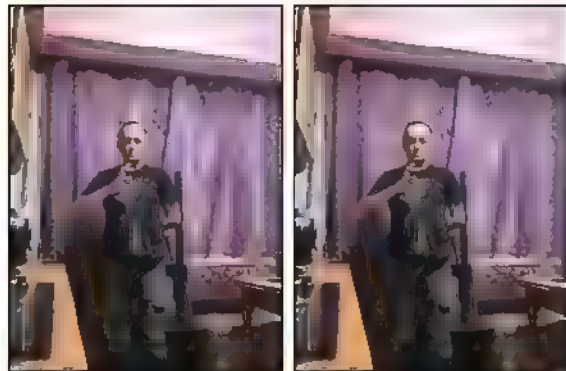
дубликат



# **Пайплайн сбора и разметки данных** уровни недобросовестных исполнителей

0 lvl

дубликат



5 lvl

шаблон





# Пайплайн сбора и разметки данных уровни недобросовестных исполнителей

0 lvl

дубликат



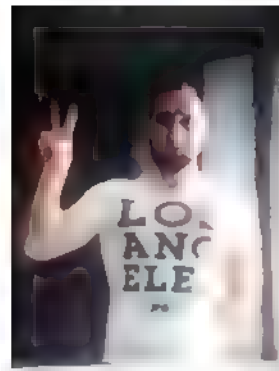
5 lvl

шаблон



20 lvl

шаблон

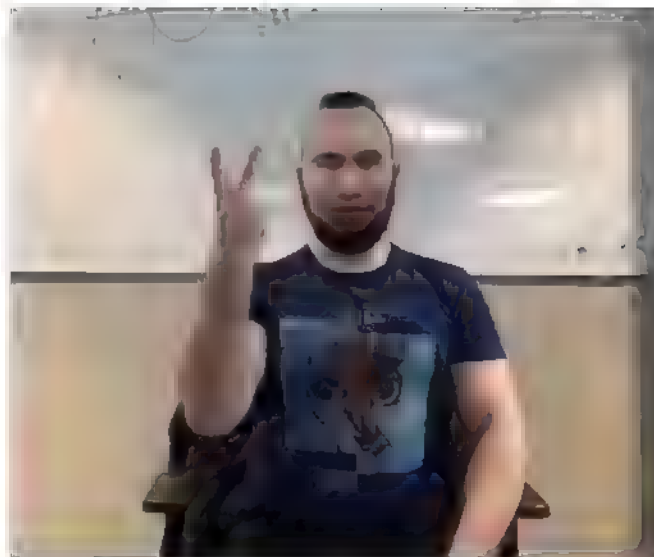
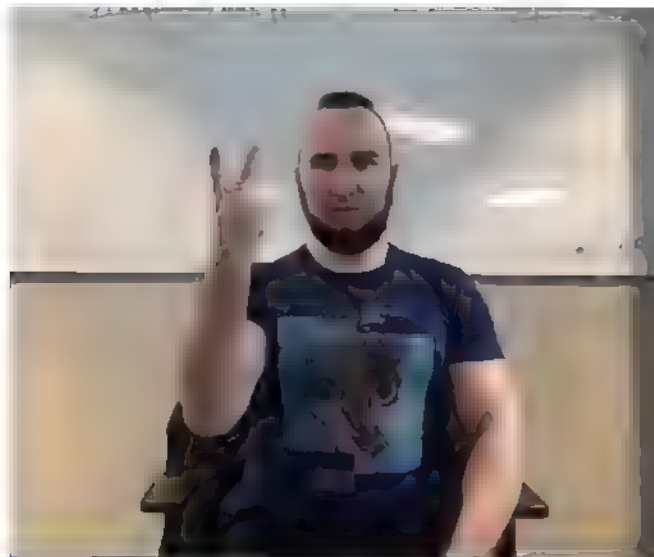


с обработкой

# Пайплайн сбора и разметки данных уровни недобросовестных исполнителей

80 lvl

дипфейки



просто fun to  
remember



## Пайплайн сбора и разметки данных

Yandex.Toloka  
валидация

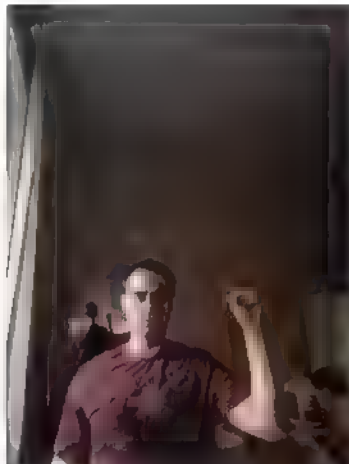
# Пайплайн сбора и разметки данных валидация

Жест кулак (рука сжата в кулак) с расстояния 3-4 метра от камеры

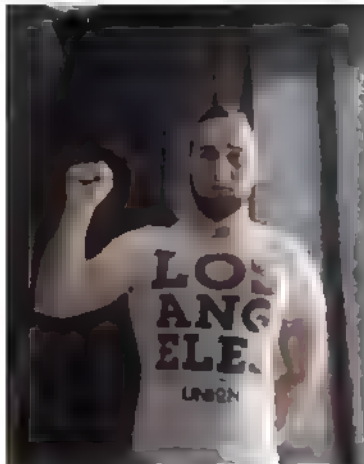
Выберите подходящий вариант ответа.

- 1 Правильно
- 2 Неправильно
- 3 Фото повернуто
- 4 Стороннее фото
- 5 Не загрузилось

Фото для оценки



Шаблон (эталонный вариант жеста)



Задача:

определить, **верно ли выполнен жест**  
на фото в соответствии с шаблоном

Критерии выбора:

жест показан **верно**

фото / жест **не соответствует критериям**

изображение **повернуто**

**стороннее фото** из интернета

**не загрузилось**

# Пайплайн сбора и разметки данных

валидация

пул  
обучение

100% лучших пользователей  
(отобраны Я.Толокой)

10 заданий с подсказкой

навык обучения  
% верных ответов

0 \$

# Пайплайн сбора и разметки данных

валидация

пул  
обучение

 пул  
экзамен

100% лучших пользователей  
(отобраны Я.Толокой)

пользователи с навыком  
обучения 70% и выше

10 заданий с подсказкой

10 заданий без подсказки

навык обучения  
% верных ответов

навык экзамена  
% верных ответов

0 \$

0 \$

# Пайплайн сбора и разметки данных

## валидация

пул  
обучение

 пул  
экзамен

пул  
основной

100% лучших пользователей  
(отобраны Я.Толокой)

пользователи с навыком  
обучения 70% и выше

пользователи с навыком  
экзамена 70% и выше

10 заданий с подсказкой

10 заданий без подсказки

10% контрольных заданий

навык обучения  
% верных ответов

навык экзамена  
% верных ответов

3 разметки на каждое  
изображение

0 \$

0 \$

0.01 \$ / 0.06 \$ / 0.09 \$



# Пайплайн сбора и разметки данных

1. Сбор данных

2. Предобработка

3. Разметка

4. Валидация

5. Обучение модели

Yandex.Toloka &  
ABC Elementary

разметка  
боксами

валидация  
разметки

Yandex.Toloka &  
ABC Elementary



# Пайплайн сбора и разметки данных

## разметка боксами

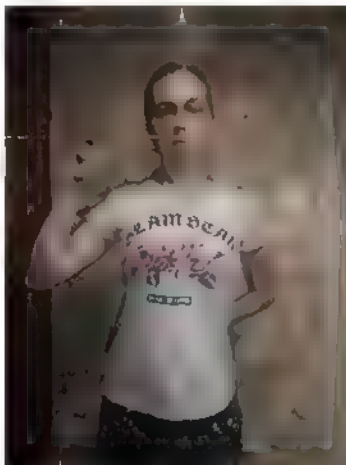
### Условия:

- ЕСЛИ КИСТЬ ОБРЕЗАНА - ОТМЕЧАТЬ НЕ НУЖНО!
- Не забывайте указывать метки Жест и Не жест
- Сторонние фото, отсутствие кистей - помечаем чекбоксом

Жест кулак (рука сжата в кулак) с расстояния 3-4 метра от камеры

Нет кистей / стороннее фото / повернуто

Фото на разметку



1 Жест 2 Не жест

Шаблон (эталонный вариант жеста)



### Задача:

выделить **кисть руки с жестом** красным  
прямоугольником, а **без жеста** – зеленым

### Условия:

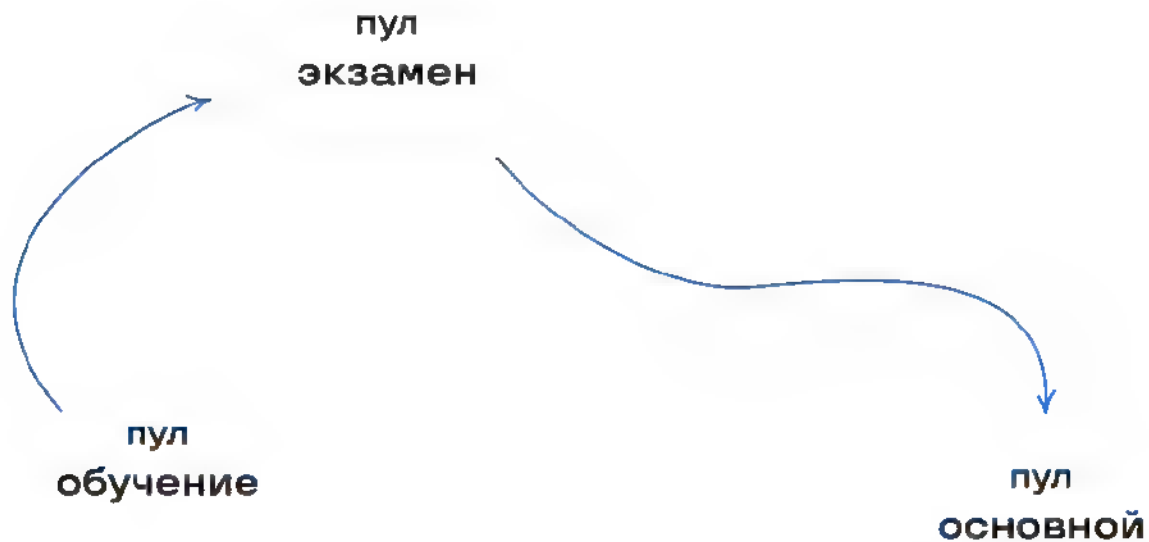


не выделять кисть, если она **обрезана**

каждому из боксов необходимо  
сопоставить метку – “**жест**” или “**не жест**”

# Пайплайн сбора и разметки данных

## разметка боксами



# Пайплайн сбора и разметки данных

## валидация разметки

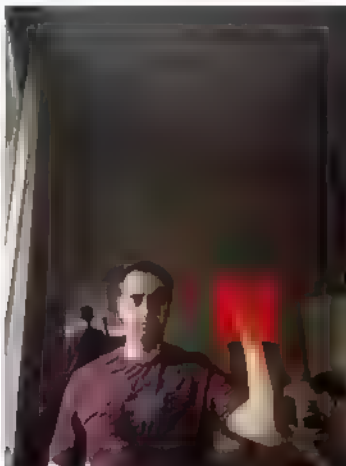
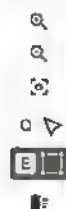
### Условия:

- ЕСЛИ КИСТЬ ОБРЕЗАНА - ОТМЕЧАТЬ НЕ НУЖНО!
- Не забывайте указывать метки Жест и Не жест
- Сторонние фото, отсутствие кистей - помечаем чекбоксом

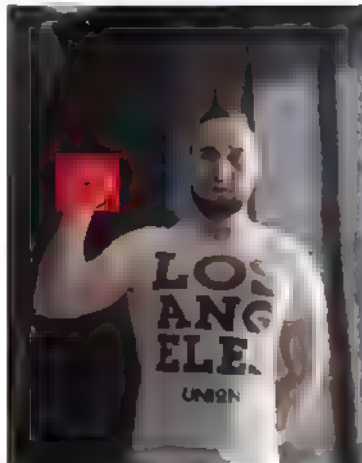
Жест кулак (рука сжата в кулак) с расстояния 3-4 метра от камеры

Нет кистей / стороннее фото / повернуто

Фото на разметку



Шаблон (эталонный вариант жеста)



### Задача:

определить, **верно ли выделены кисти** рук на фото

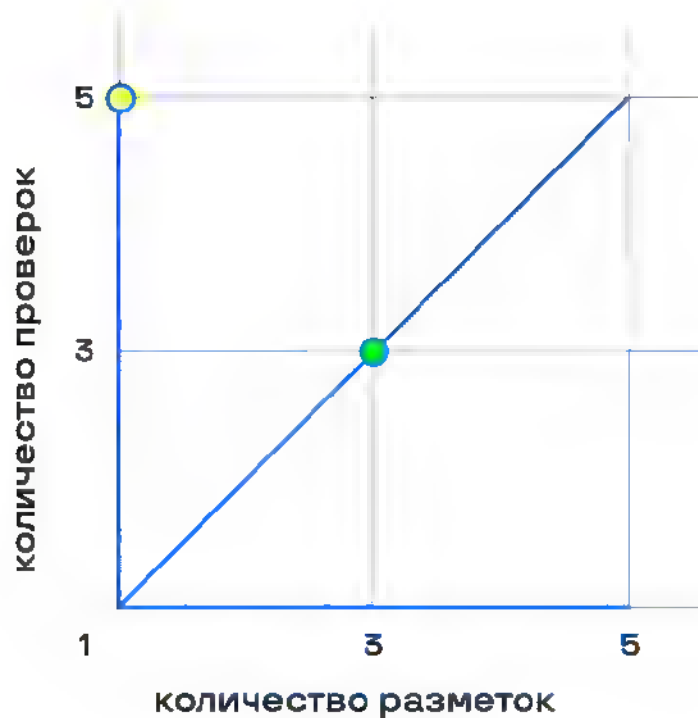
### Критерии выбора:

все **кисти четко очерчены** боксами

каждый из боксов имеет правильную метку – **“жест”** или **“не жест”**

# Пайплайн сбора и разметки данных

разметка боксами и валидация разметки



↑ воздействие разметчиков

↑ качество

↓ уверенность в разметке

↑ стоимость

# Пайплайн сбора и разметки данных

сбор изображений



удаляем дубликаты  
& фото с низким  
разрешением



валидация



разметка  
боксами



валидация  
разметки



HaGRID



## Стоимость пайплайна as is

каждое задание оценивалось в **один цент** ( $s = 0.01\$$ )

	n	по этапам	1 изображение
сбор изображений	878.910	$n \times s \approx 9.000\$$	0.01\$
валидация	830.679	$n \times s \times 3 \approx 25.000\$$	$s \times 3 \approx 0.03\$$
разметка	560.230	$n \times s \times 5 \approx 28.000\$$	$s \times 5 \approx 0.05\$$
валидация разметки	560.230	$n \times s \times 5 \times 5 \approx 140.000\$$	$s \times 5 \times 5 \approx 0.25\$$
классификация	552.992	$n \times s \times 3 \approx 17.000\$$	$s \times 3 \approx 0.03\$$
итого	552.992	?	0.37\$



## Стоимость пайплайна as is



220.000\$

BMW 8 серии Cabrio



или

90 x



или

440 x

SberPortal





**дорого  
&  
некачественно**

много лет мне  
заняло, чтобы  
эта idea create

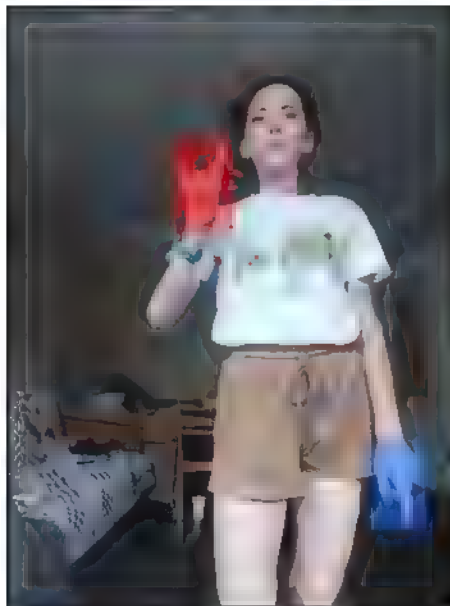


**валидации разметки**  
может, ее автоматизировать?

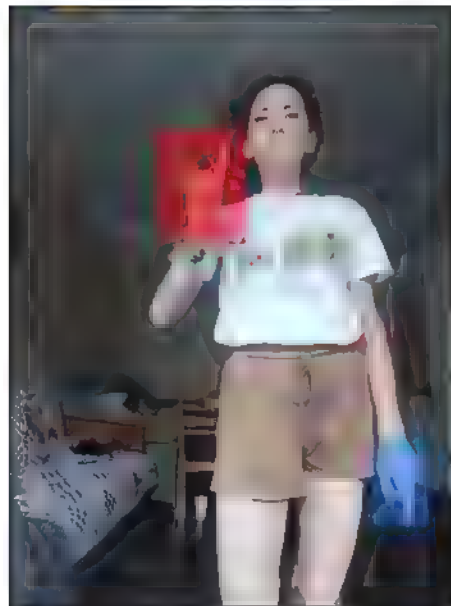
# Автоматизация валидации разметки

а какая бывает разметка?

идеальная



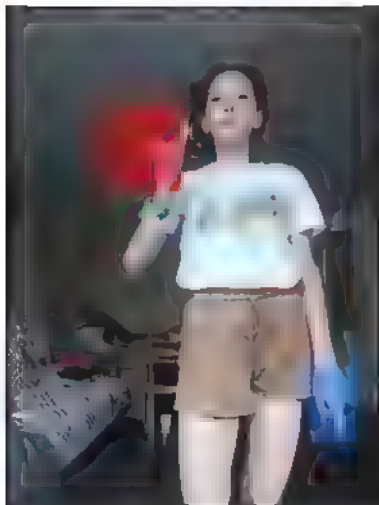
хорошая



# Автоматизация валидации разметки

а какая бывает разметка?

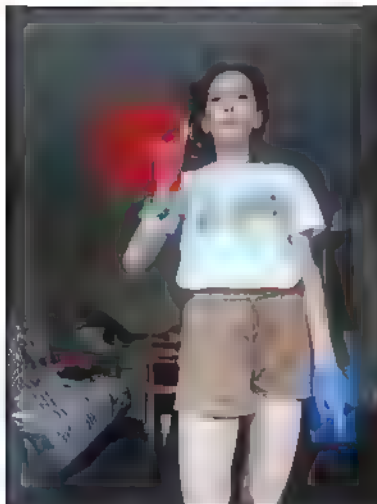
бокс не там,  
где нужно :(



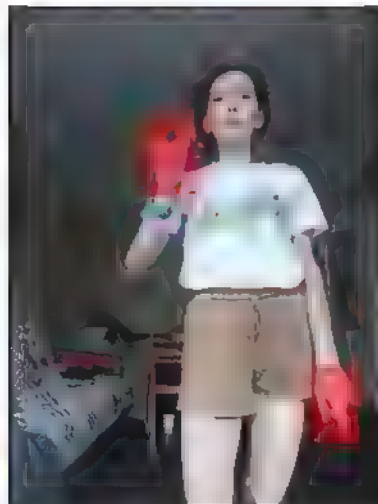
# Автоматизация валидации разметки

а какая бывает разметка?

бокс не там,  
где нужно :(



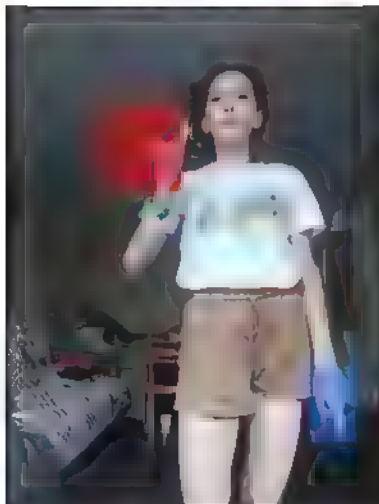
не та метка :(



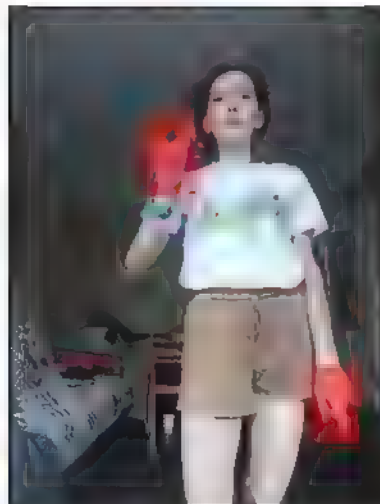
# Автоматизация валидации разметки

а какая бывает разметка?

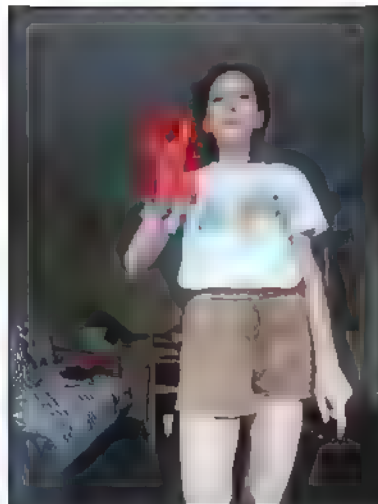
бокс не там,  
где нужно :(



не та метка :(



забыли бокс :(

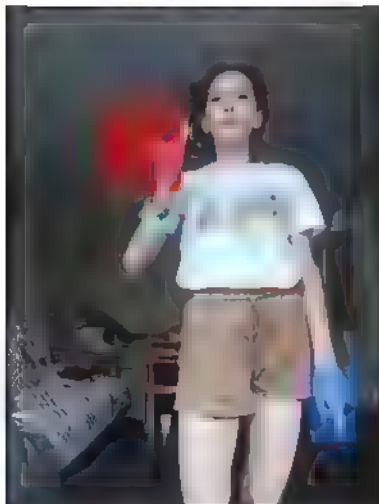


# Автоматизация валидации разметки

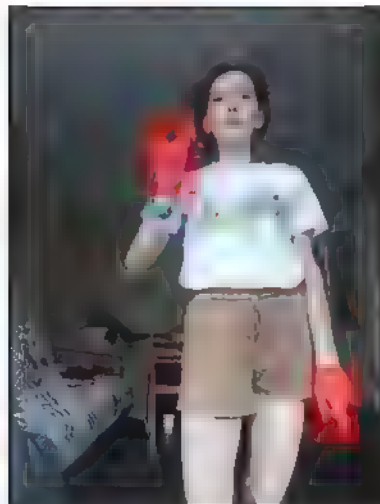
а какая бывает разметка?



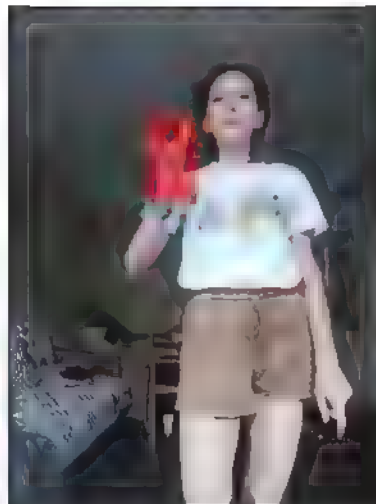
бокс не там,  
где нужно :(



не та метка :(



забыли бокс :(



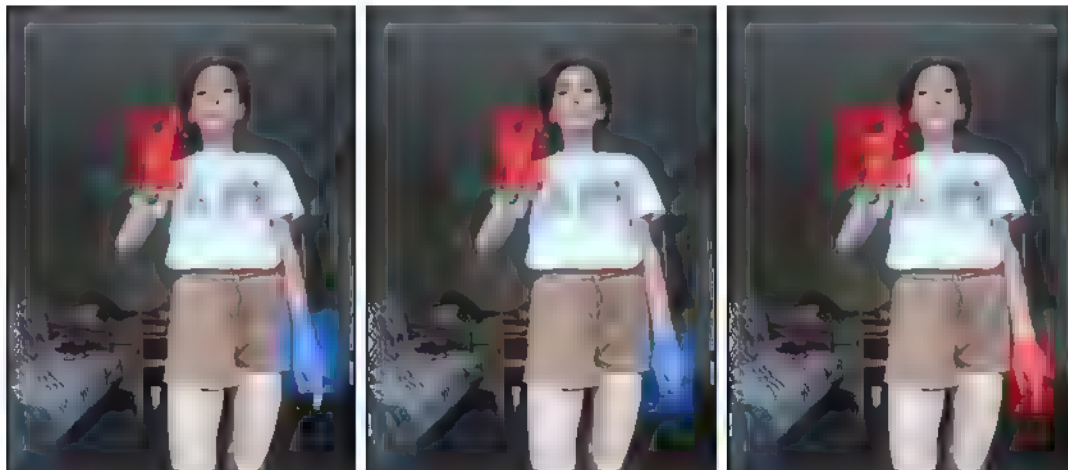
нуу..





# Агрегация разметки

3 разметки



## Агрегация разметки



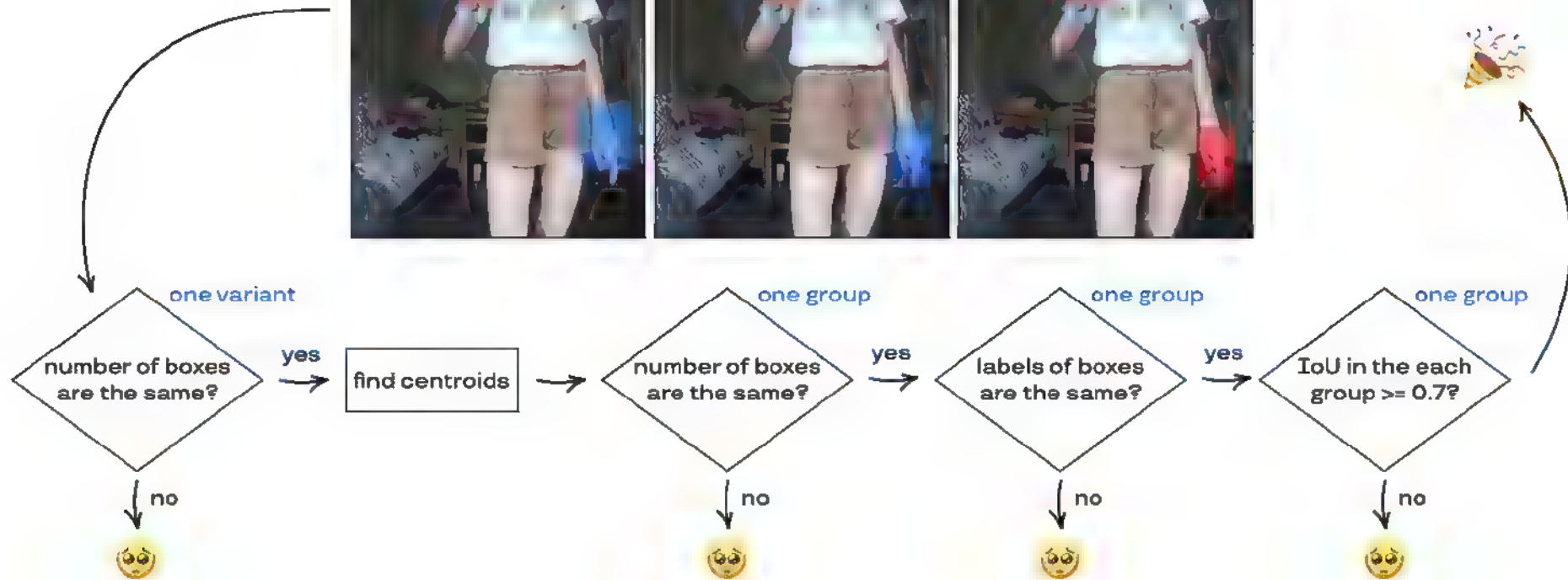
3 разметки



hard

# Агрегация разметки

hard-подход



# Агрегация разметки

hard-подход

во всех разметках по 2 бокса 😎



one variant

number of boxes  
are the same?

yes

no



# Агрегация разметки

hard-подход

в каждой группе по 3 бокса 😊



one variant

number of boxes  
are the same?

yes

find centroids

one group

number of boxes  
are the same?

yes

no



# Агрегация разметки

hard-подход

в одной разметке неверно указана метка 🤔



one variant

number of boxes  
are the same?

yes

find centroids

one group

number of boxes  
are the same?

yes

one group

labels of boxes  
are the same?

yes

no



no



no



# Агрегация разметки

hard-подход

IoU в каждой из групп больше 70% 😊



one variant

number of boxes  
are the same?

yes

find centroids

one group

number of boxes  
are the same?

yes

one group

labels of boxes  
are the same?

yes

IoU in the each  
group  $\geq 0.7$ ?

one group

no



no



no



no



## Агрегация разметки

3 разметки



hard

успех?



разметка



# Агрегация разметки



# Агрегация разметки



# Агрегация разметки

4 разметки



hard

не успех?

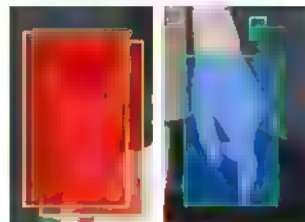


## Агрегация разметки

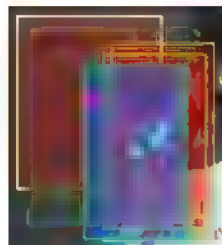


# Агрегация разметки

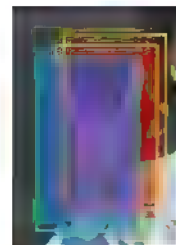
soft-подход



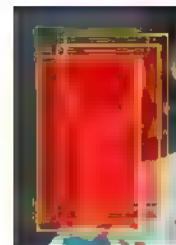
remove boxes close to dots & duplicates



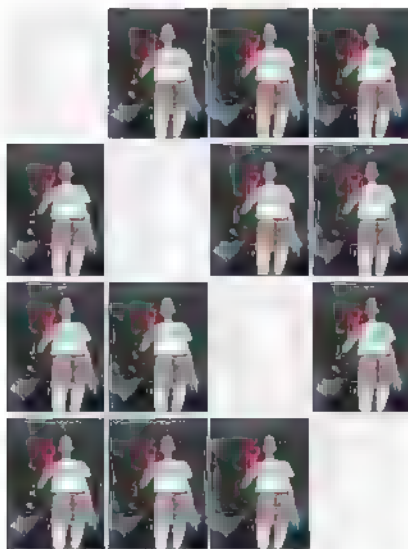
find centroids & remove "outliers"



replace incorrect labels



add missing boxes



iteratively drop k markup variants

hard aggregation

yes

no



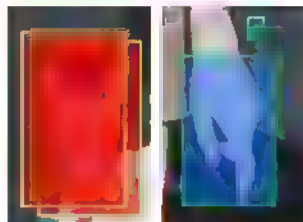
# Агрегация разметки

soft-подход



# Агрегация разметки

soft-подход



remove boxes close  
to dots & duplicates



# Агрегация разметки

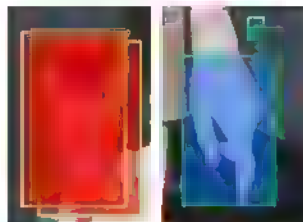
soft-подход



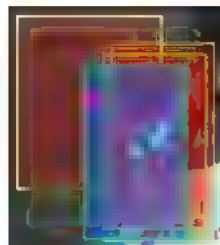


# Агрегация разметки

soft-подход



remove boxes close  
to dots & duplicates

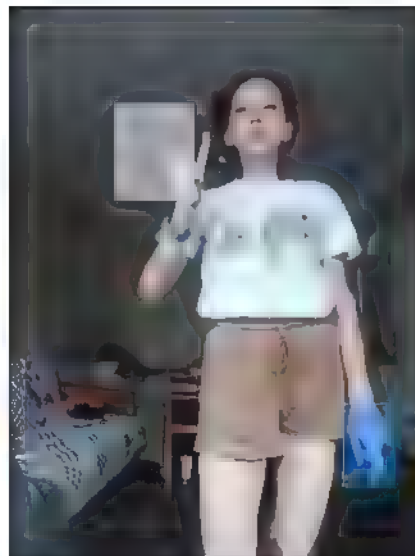
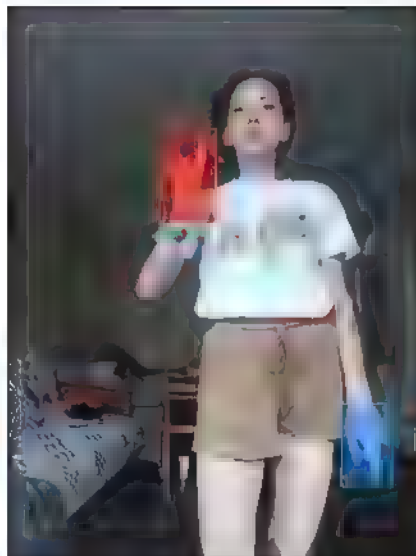
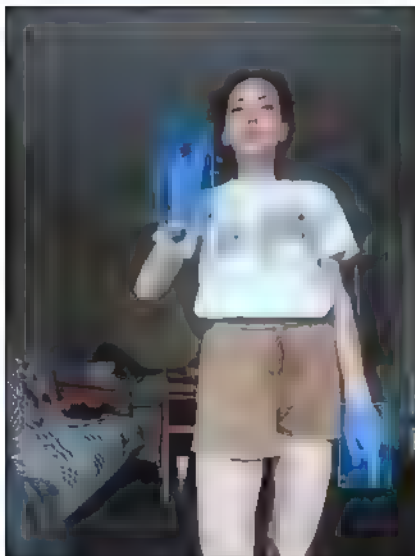


find centroids &  
remove "outliers"



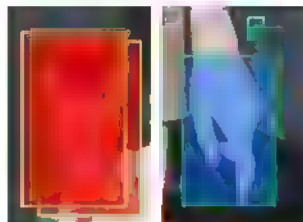
# Агрегация разметки

soft-подход

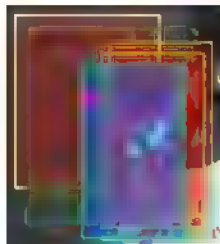


# Агрегация разметки

soft-подход



remove boxes close  
to dots & duplicates



find centroids &  
remove "outliers"



replace incorrect  
labels



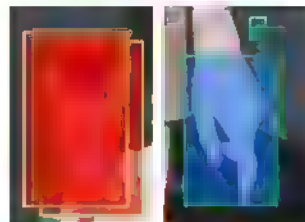
# Агрегация разметки

soft-подход

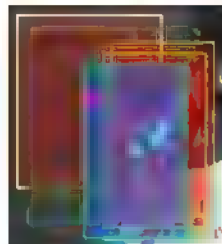


# Агрегация разметки

soft-подход



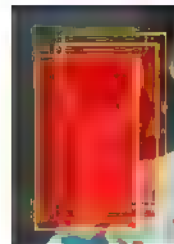
remove boxes close  
to dots & duplicates



find centroids &  
remove "outliers"



replace incorrect  
labels



add missing boxes



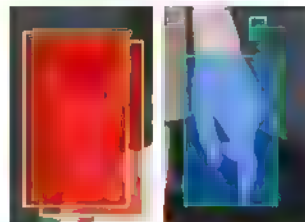
# Агрегация разметки

soft-подход

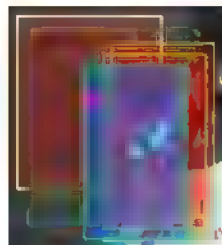


# Агрегация разметки

soft-подход



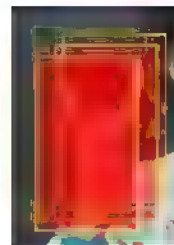
remove boxes close to dots & duplicates



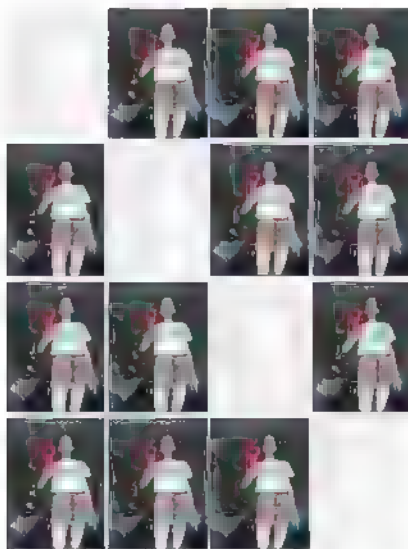
find centroids & remove "outliers"



replace incorrect labels



add missing boxes



iteratively drop k markup variants

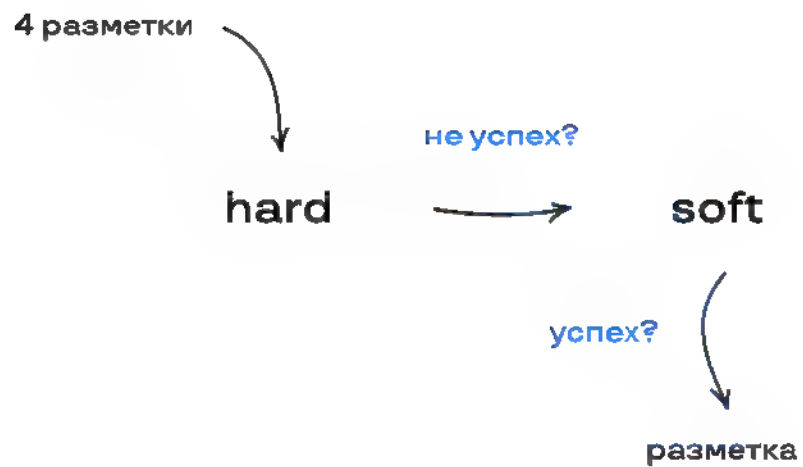
hard aggregation

yes

no

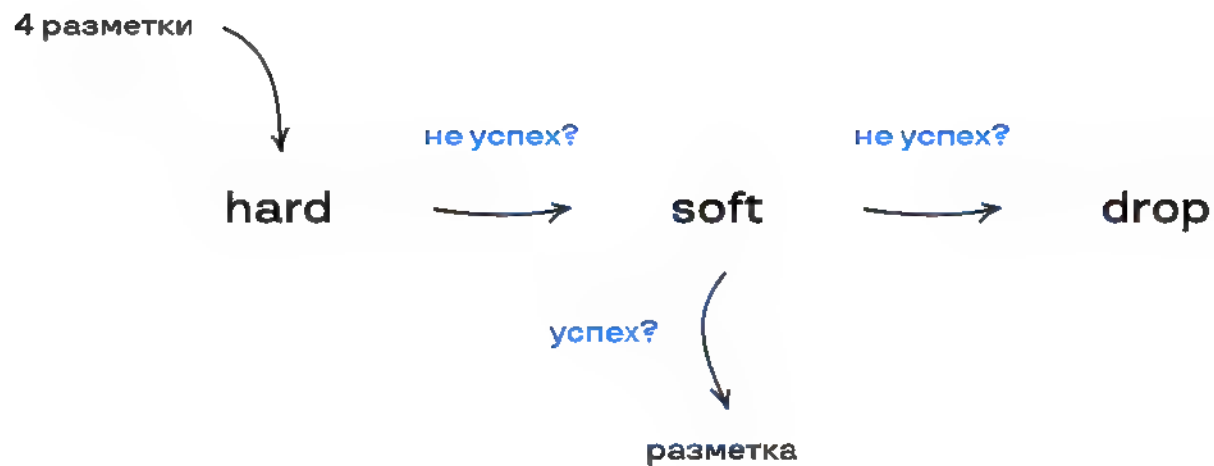


# Агрегация разметки



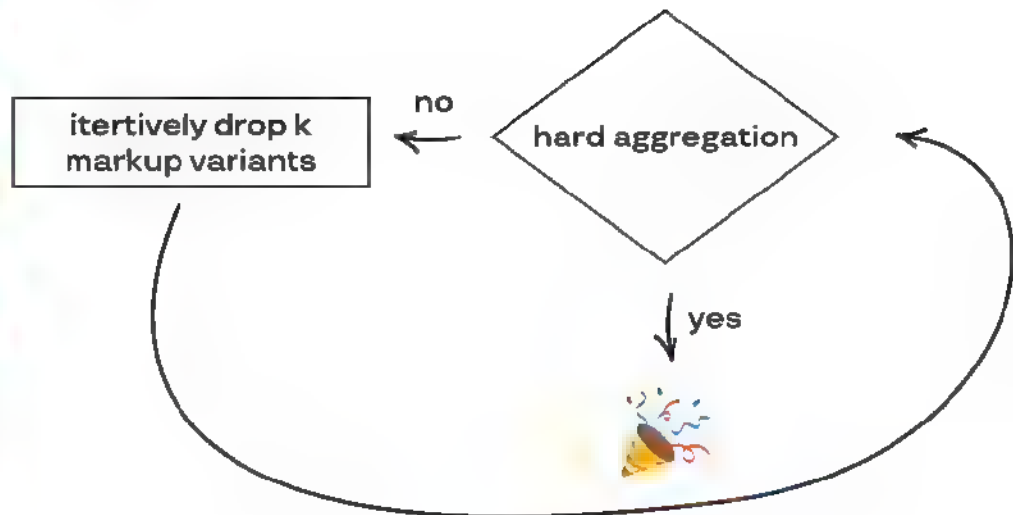


# Агрегация разметки

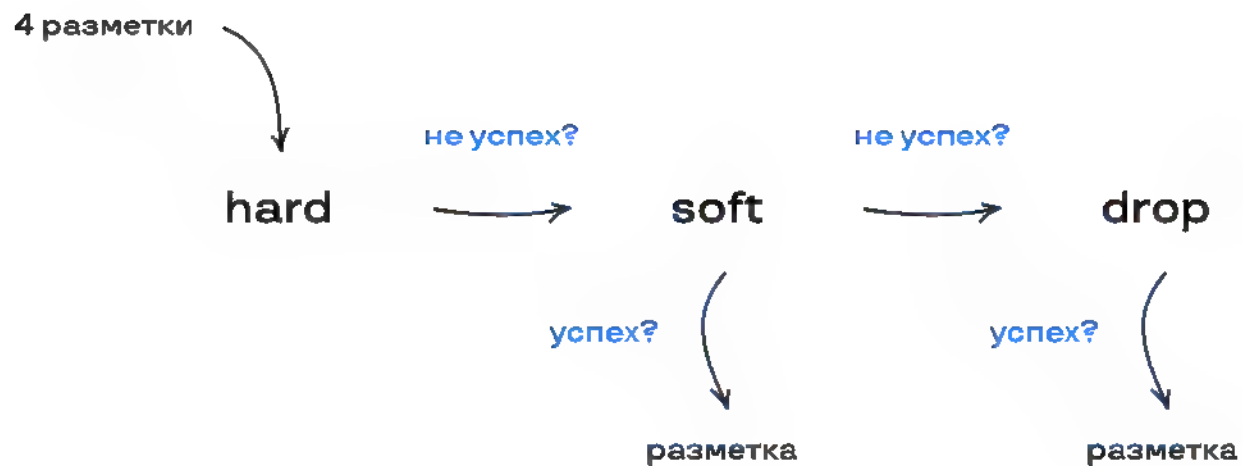


# Агрегация разметки

## drop-подход



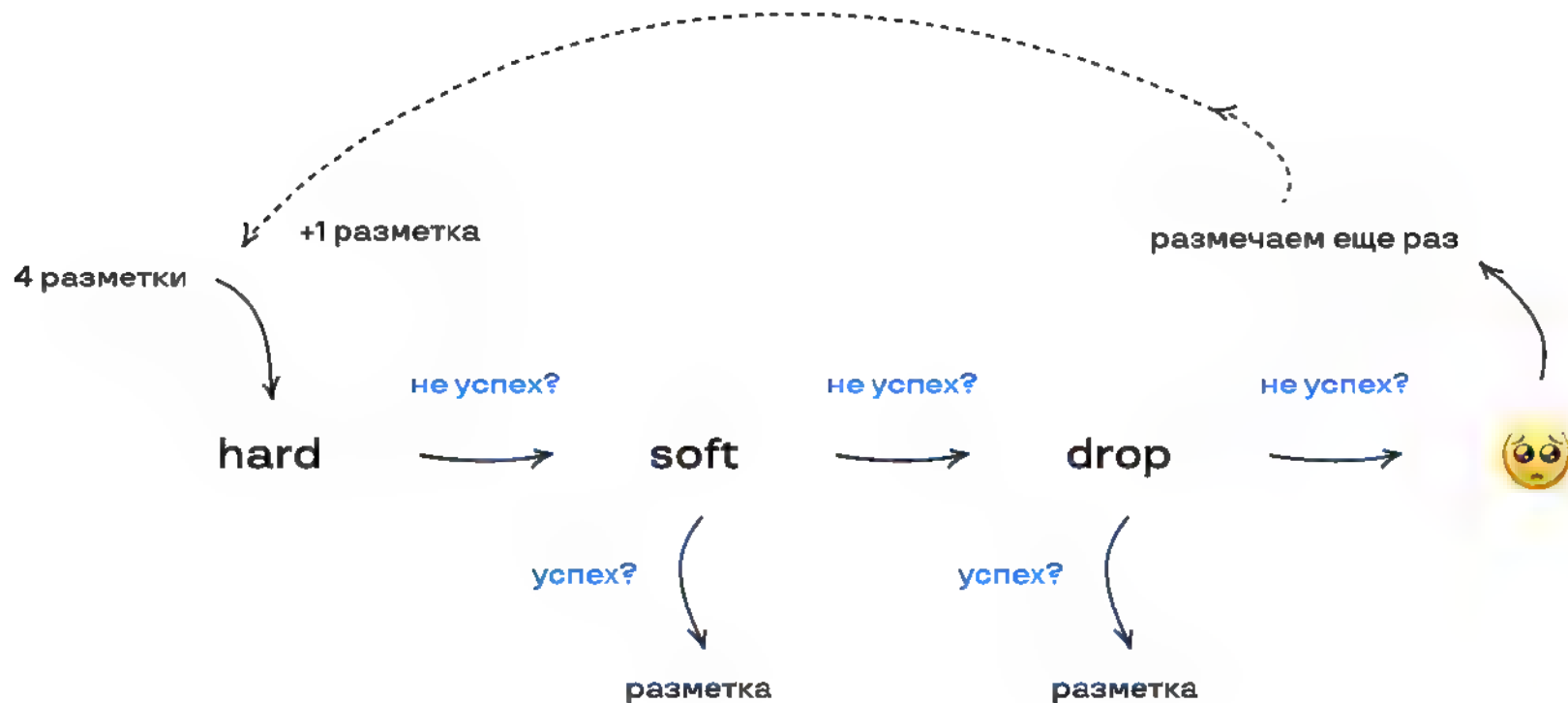
# Агрегация разметки



# Агрегация разметки



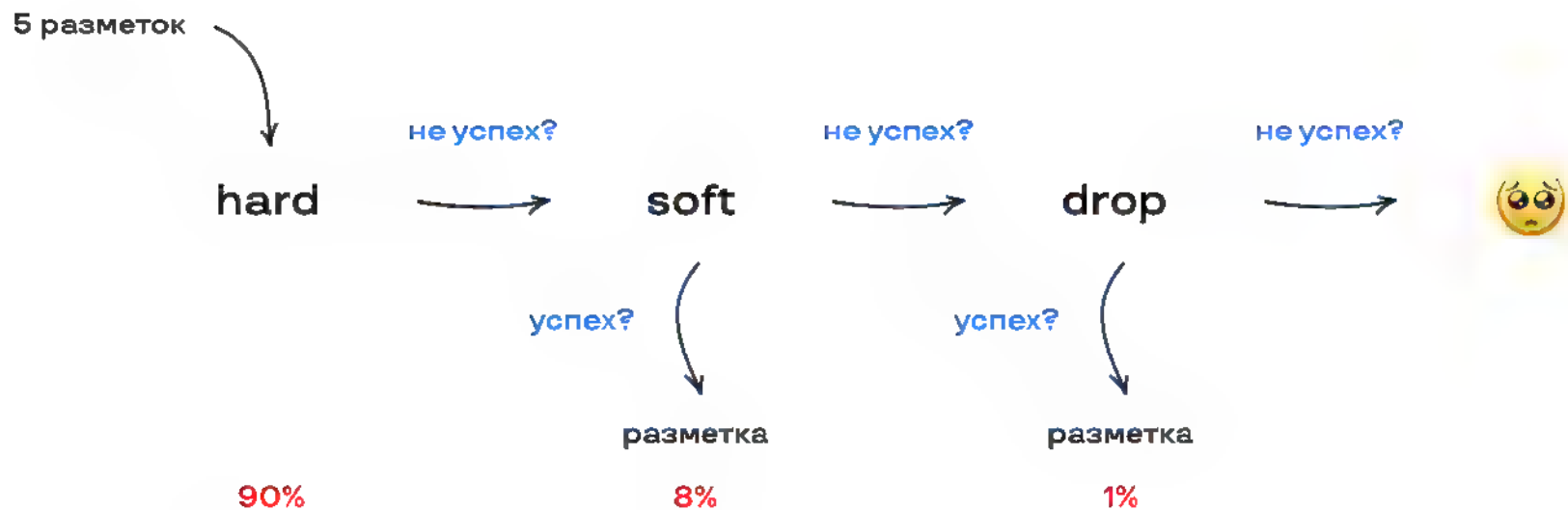
# Агрегация разметки



# Агрегация разметки



# Агрегация разметки



# Пайплайн сбора и разметки данных

сбор изображений



удаляем дубликаты  
& фото с низким  
разрешением



валидация



разметка  
боксами



валидация  
агрегация  
разметки



HaGRID





## Стоимость пайплайна

каждое задание оценивалось в **один цент** ( $s = 0.01\$$ )

	n	по этапам	1 изображение
сбор изображений	878.910	$n \times s \approx 9.000\$$	0.01\$
валидация	830.679	$n \times s \times 3 \approx 25.000\$$	$s \times 3 \approx 0.03\$$
разметка	560.230	$n \times s \times 5 \approx 28.000\$$	$s \times 5 \approx 0.05\$$
агрегация разметки	560.230	140.000\$ → 0\$	0.25\$ → 0\$
итого	552.992	220.000\$ → 62.000\$	0.37\$ → 0.09\$

# Публикация в open-source

сбор изображений



удаляем дубликаты  
& фото с низким разрешением



валидация



разметка  
боксами



агрегация  
разметки



HaGRID

а может, в open-source?



# Публикация в open-source

что стоит сделать для публикации в open-source?

## Отсеять изображения

- 👶 с детьми в кадре
- 👮 с людьми без одежды (да-да, и такое бывает)
- 💧 с вотермарками
- 🚫 с запрещенной символикой

это очень  
important



# Публикация в open-source

сбор изображений



удаляем дубликаты  
& фото с низким  
разрешением



валидация



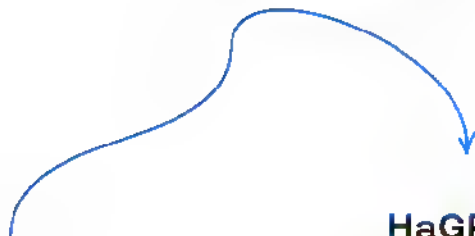
разметка  
боксами



агрегация  
разметки



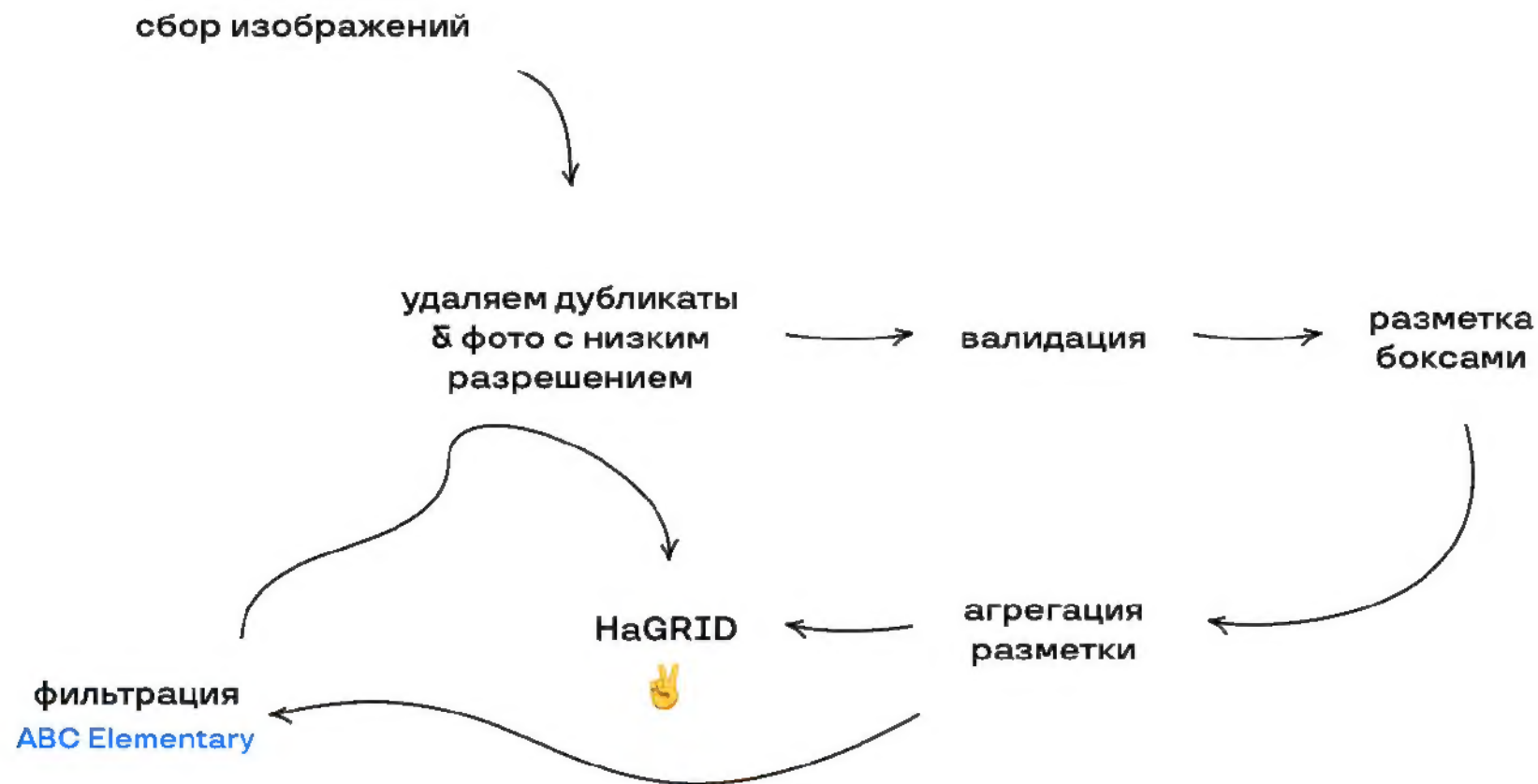
HaGRID



фильтрация



# Публикация в open-source



# Автоматизированный пайплайн

```
> python mining -p gestures -d image -c toloka -t <list_of_gestures> -overlap 1000
```



Название ↕	Приоритет ↕	Прогресс	Статус ↕	Запущен ↕
CALL [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 8005* / 10000	Идёт разметка...	18 октября 2022 г., 9:22
LIKE [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 6143* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
DISLIKE [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 6134* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
FIST [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 6785* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
STOP [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 6645* / 9000	Идёт разметка...	18 октября 2022 г., 22:56
STOP_INVERTED [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 1765* / 2000	Идёт разметка...	13 октября 2022 г., 13:20
THREE [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 6273* / 8000	Идёт разметка...	18 октября 2022 г., 9:21
TWO_UP [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 712* / 1000	Идёт разметка...	18 октября 2022 г., 21:58
THREE2 [RNDCV] [Alexander Kapitanov]	0	<div><div></div></div> 681* / 1000	Идёт разметка...	18 октября 2022 г., 22:22

# Автоматизированный пайплайн

какие еще данные собирали

skin segmentation

teeth segmentation

image to text

text to image

side by side

portrait segmentation

face detection

arxiv



github



<https://arxiv.org/abs/2206.08219>

<https://github.com/hukenovs/hagrid>

Александр Капитанов  
Team Lead CV RnD  
[@hukenovs](#)

Карина Кванчиани  
CV Engineer  
[@karinakv](#)



оценить доклад

